

# Knowledge Discovery and Data Mining

Unit # 15

## Acknowledgement

- Most of the slides in this presentation are taken from course slides provided by
  - Han and Kimber (Data Mining Concepts and Techniques) and
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

# Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

## Market-Basket transactions

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Example of Association Rules

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$   
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$   
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

Implication means co-occurrence, not causality!

Sajjad Haider

Fall 2012

3

# Definition: Frequent Itemset

- Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- Support count ( $\sigma$ )**
  - Frequency of occurrence of an itemset
  - E.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- Support**
  - Fraction of transactions that contain an itemset
  - E.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Definition: Frequent Itemset

## Definition: Association Rule

- Association Rule

- An implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets
- Example:  
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Rule Evaluation Metrics

- Support ( $s$ )
  - ◆ Fraction of transactions that contain both  $X$  and  $Y$
- Confidence ( $c$ )
  - ◆ Measures how often items in  $Y$  appear in transactions that contain  $X$

Example:

$$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

## Association Rule Mining Task

- Given a set of transactions  $T$ , the goal of association rule mining is to find all rules having
  - support  $\geq \text{minsup}$  threshold
  - confidence  $\geq \text{minconf}$  threshold
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

$\Rightarrow$  **Computationally prohibitive!**

## Apriori

- Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties*.
- Two steps process
  - Join
  - Prune

Sajjad Haider

Fall 2012

7

## Mining Association Rules

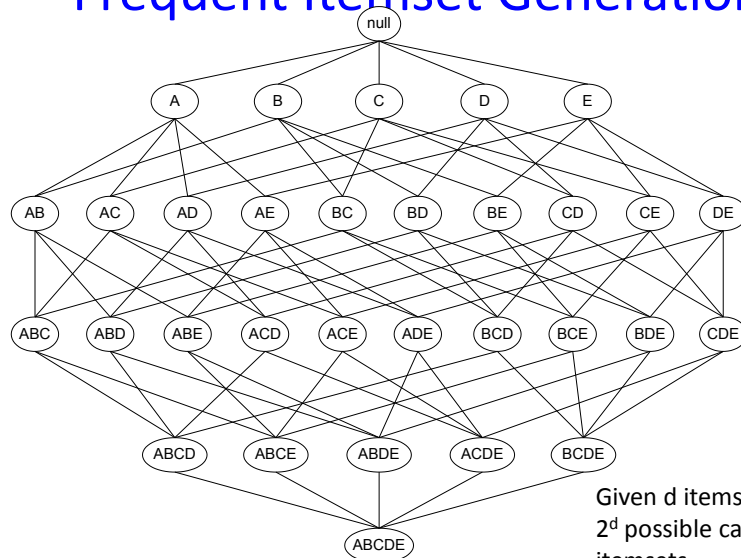
- Two-step approach:
  1. **Frequent Itemset Generation**
    - Generate all itemsets whose support  $\geq$  minsup
  2. **Rule Generation**
    - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
- Frequent itemset generation is still computationally expensive

Sajjad Haider

Fall 2012

8

## Frequent Itemset Generation



Given  $d$  items, there are  $2^d$  possible candidate itemsets

Sajjad Haider

Fall 2012

9

## Reducing Number of Candidates

- **Apriori principle:**
  - If an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

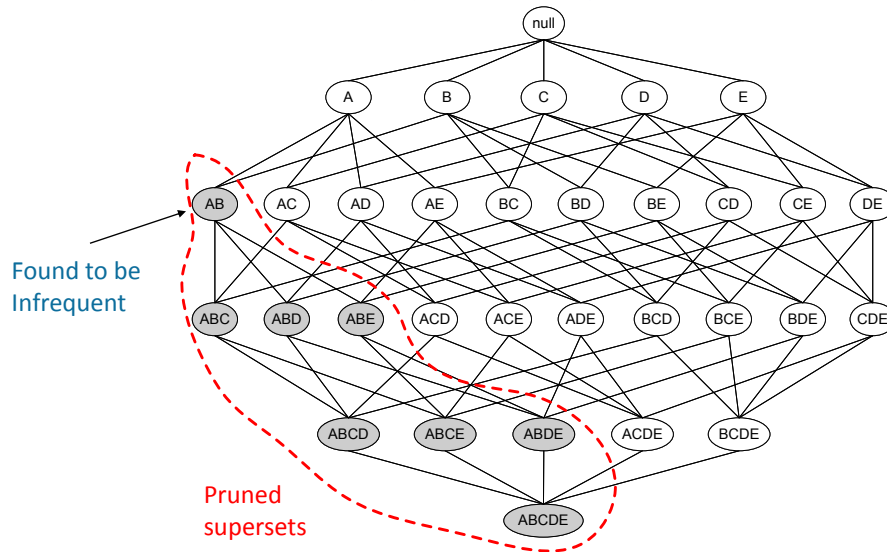
- Support of an itemset never exceeds the support of its subsets
- This is known as the **anti-monotone** property of support

Sajjad Haider

Fall 2012

10

## Illustrating Apriori Principle



Sajjad Haider

Fall 2012

11

## Illustrating Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Itemset	Count
{Bread, Milk}	3
{Bread, Beer}	2
{Bread, Diaper}	3
{Milk, Beer}	2
{Milk, Diaper}	3
{Beer, Diaper}	3

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Minimum Support = 3

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3 = 41$   
 With support-based pruning,  
 $6 + 6 + 1 = 13$

Itemset	Count
{Bread, Milk, Diaper}	3

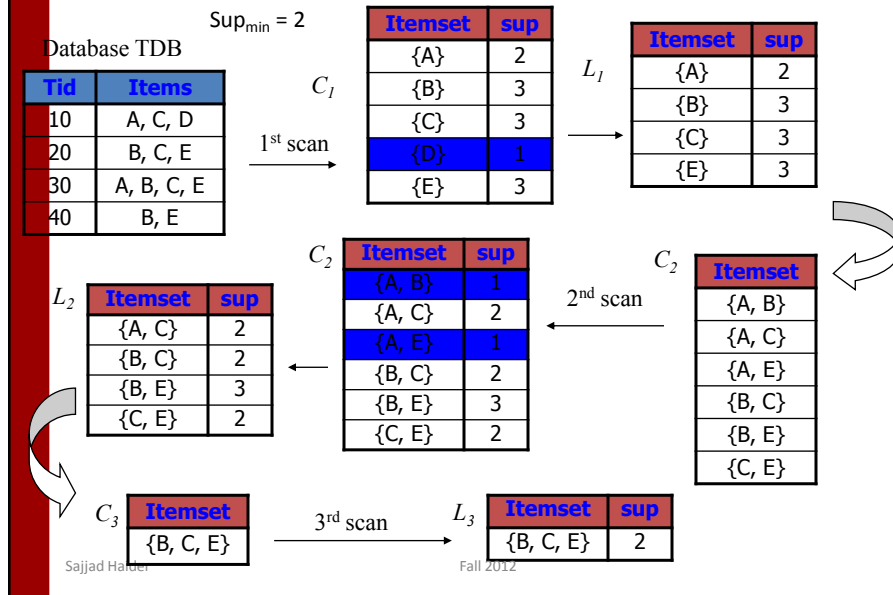
Triplets (3-itemsets)

Sajjad Haider

Fall 2012

12

## The Apriori Algorithm—An Example



## Apriori Algorithm

- Method:
  - Let  $k=1$
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length  $(k+1)$  candidate itemsets from length  $k$  frequent itemsets
    - Prune candidate itemsets containing subsets of length  $k$  that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

## Factors Affecting Complexity

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase
- Size of database
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions

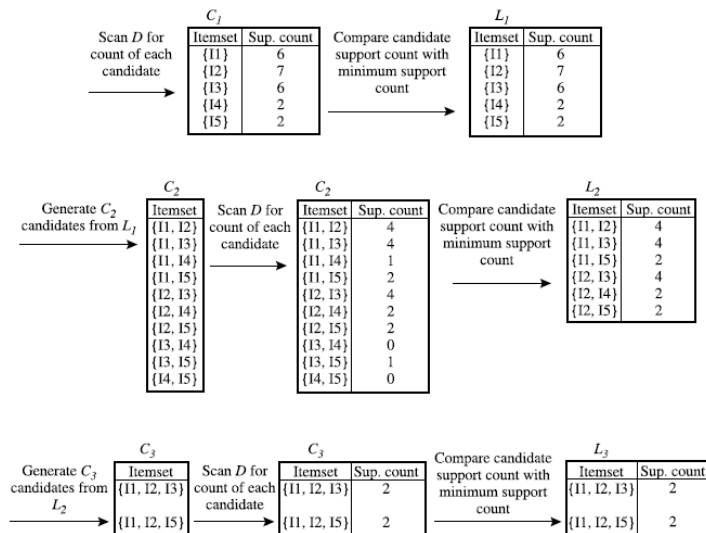
## Example

<i>TID</i>	<i>List of item IDs</i>
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Minimum Support = 2



## Example (Cont'd)



Sajjad Hai

17

## Rule Generation

- Given a frequent itemset  $L$ , find all non-empty subsets  $f \subset L$  such that  $f \rightarrow L - f$  satisfies the minimum confidence requirement
  - If  $\{A, B, C, D\}$  is a frequent itemset, candidate rules:
 

$ABC \rightarrow D,$	$ABD \rightarrow C,$	$ACD \rightarrow B,$	$BCD \rightarrow A,$
$A \rightarrow BCD,$	$B \rightarrow ACD,$	$C \rightarrow ABD,$	$D \rightarrow ABC$
$AB \rightarrow CD,$	$AC \rightarrow BD,$	$AD \rightarrow BC,$	$BC \rightarrow AD,$
$BD \rightarrow AC,$	$CD \rightarrow AB,$		
- If  $|L| = k$ , then there are  $2^k - 2$  candidate association rules (ignoring  $L \rightarrow \emptyset$  and  $\emptyset \rightarrow L$ )

Sajjad Haider

Fall 2012

18

## Rule Generation

- How to efficiently generate rules from frequent itemsets?

- In general, confidence does not have an anti-monotone property  
 $c(ABC \rightarrow D)$  can be larger or smaller than  $c(AB \rightarrow D)$

- But confidence of rules generated from the same itemset has an anti-monotone property

- e.g.,  $L = \{A, B, C, D\}$ :

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

- Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

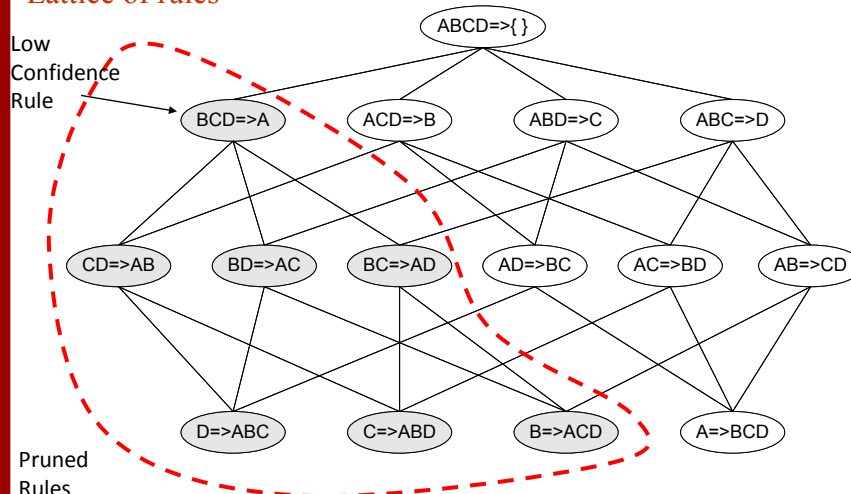
Sajjad Haider

Fall 2012

19

## Rule Generation for Apriori Algorithm

### Lattice of rules



Sajjad Haider

Fall 2012

20

## Rule Generation Example

- Suppose the data contains the frequent itemset  $I = \{I1, I2, I5\}$ . What are the association rules that can be generated from  $I$ ?
- The nonempty subsets of  $I$  are  $\{I1, I2\}$ ,  $\{I1, I5\}$ ,  $\{I2, I5\}$ ,  $\{I1\}$ ,  $\{I2\}$ , and  $\{I5\}$ .
- If the minimum confidence threshold is, say, 70%, then only the second, third, and last rules below are output

$I1 \wedge I2 \Rightarrow I5$ ,	$confidence = 2/4 = 50\%$
$I1 \wedge I5 \Rightarrow I2$ ,	$confidence = 2/2 = 100\%$
$I2 \wedge I5 \Rightarrow I1$ ,	$confidence = 2/2 = 100\%$
$I1 \Rightarrow I2 \wedge I5$ ,	$confidence = 2/6 = 33\%$
$I2 \Rightarrow I1 \wedge I5$ ,	$confidence = 2/7 = 29\%$
$I5 \Rightarrow I1 \wedge I2$ ,	$confidence = 2/2 = 100\%$

## Effect of Support Distribution

- How to set the appropriate *minsup* threshold?
  - If *minsup* is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)
  - If *minsup* is set too low, it is computationally expensive and the number of itemsets is very large

## Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if  $\{A,B,C\} \rightarrow \{D\}$  and  $\{A,B\} \rightarrow \{D\}$  have same support & confidence
- Interestingness measures can be used to prune/rank the derived patterns
- In the original formulation of association rules, support & confidence are the only measures used

Sajjad Haider

Fall 2012

23

## Computing Interestingness Measure

- Given a rule  $X \rightarrow Y$ , information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for  $X \rightarrow Y$

	Y	$\overline{Y}$	
X	$f_{11}$	$f_{10}$	$f_{1+}$
$\overline{X}$	$f_{01}$	$f_{00}$	$f_{0+}$
	$f_{+1}$	$f_{+0}$	$ T $

$f_{11}$ : support of X and Y

$f_{10}$ : support of X and  $\overline{Y}$

$f_{01}$ : support of  $\overline{X}$  and Y

$f_{00}$ : support of  $\overline{X}$  and  $\overline{Y}$

Sajjad Haider

Fall 2012

24

## Drawback of Confidence

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

Association Rule: Tea  $\rightarrow$  Coffee

Confidence =  $P(\text{Coffee}|\text{Tea}) = 0.75$

but  $P(\text{Coffee}) = 0.9$

$\Rightarrow$  Although confidence is high, rule is misleading

$\Rightarrow P(\text{Coffee}|\overline{\text{Tea}}) = 0.9375$

## Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)
  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \wedge B) = P(S) \times P(B) \Rightarrow$  Statistical independence
  - $P(S \wedge B) > P(S) \times P(B) \Rightarrow$  Positively correlated
  - $P(S \wedge B) < P(S) \times P(B) \Rightarrow$  Negatively correlated

## Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{P(Y | X)}{P(Y)}$$

$$Interest = \frac{P(X, Y)}{P(X)P(Y)}$$

$$PS = P(X, Y) - P(X)P(Y)$$

$$\phi - coefficient\ t = \frac{P(X, Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}$$

Sajjad Haider

Fall 2012

27

## Example: Lift/Interest

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence =  $P(\text{Coffee} | \text{Tea}) = 0.75$

but  $P(\text{Coffee}) = 0.9$

⇒ Lift =  $0.75/0.9 = 0.8333$  (< 1, therefore is negatively associated)

Sajjad Haider

Fall 2012

28

	#	Measure	Formula	
There are lots of measures proposed in the literature	1	$\phi$ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$	
	2	Goodman-Kruskal's $\lambda$	$\frac{\sum_j \max_k P(A_j, B_k) - \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$	
	3	Odds ratio ( $\alpha$ )	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,B)P(\bar{A},B)}$	
	4	Yule's $Q$	$\frac{P(A,B)P(\bar{A},\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A},\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$	
Some measures are good for certain applications, but not for others	5	Yule's $Y$	$\frac{\sqrt{P(A,B)P(\bar{A},\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A},\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$	
	6	Kappa ( $\kappa$ )	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$	
	7	Mutual Information ( $M$ )	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$	
	8	J-Measure ( $J$ )	$\max \left( P(A, B) \log \left( \frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left( \frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), \right.$ $\left. P(A, B) \log \left( \frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left( \frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$	
What criteria should we use to determine whether a measure is good or bad?	9	Gini index ( $G$ )	$\max \left( P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] \right.$ $\left. - P(B)^2 - P(\bar{B})^2, \right.$ $\left. P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] \right.$ $\left. - P(A)^2 - P(\bar{A})^2 \right)$	
	10	Support ( $s$ )	$P(A, B)$	
	11	Confidence ( $c$ )	$\max(P(B A), P(A B))$	
	12	Laplace ( $L$ )	$\max \left( \frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$	
What about Apriori-style support based pruning? How does it affect these measures?	13	Conviction ( $V$ )	$\max \left( \frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{A}B)} \right)$	
	14	Interest ( $I$ )	$\frac{P(A,B)}{P(A)P(B)}$	
	15	cosine ( $IS$ )	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$	
	16	Piatetsky-Shapiro's ( $PS$ )	$P(A, B) - P(A)P(B)$	
	17	Certainty factor ( $F$ )	$\max \left( \frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$	
	18	Added Value ( $AV$ )	$\max(P(B A) - P(B), P(A B) - P(A))$	
	19	Collective strength ( $S$ )	$\frac{P(A,B) + P(\bar{A},\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$	
	20	Jaccard ( $\zeta$ )	$\frac{P(A) + P(B) - P(A,B)}{P(A,B)}$	
	21	Klogsen ( $K$ )	$\sqrt{P(\bar{A}, \bar{B}) \max(P(B A) - P(B), P(A B) - P(A))}$	

Sajjad Haider

Fall 2014

29